Odyssey 2016
21-24 Jun 2016, Bilbao, Spain

Semi-supervised On-line Speaker Diarization for Meeting Data with Incremental Maximum A-posteriori Adaptation

Giovanni Soldi[1], *Massimiliano Todisco*[1], *Héctor Delgado*[1], *Christophe Beaugeant*[2] *and Nicholas Evans*[1]

[1]EURECOM, Sophia Antipolis, France
[2] Intel, Sophia Antipolis, France,
[1] {soldi,todisco,delgado,evans}@eurecom.fr, [2]christophe.beaugeant@intel.com

Abstract

Almost all current diarization systems are off-line and ill-suited to the growing need for on-line or real-time diarization. Our previous work reported the first on-line diarization system for the most challenging speaker diarization domain involving meeting data captured with a single distant microphone (SDM). Even if results were not dissimilar to those reported for on-line diarization in less challenging domains, error rates were high and unlikely to support any practical applications. The first novel contribution in this paper relates to the investigation of a semi-supervised approach to on-line diarization whereby speaker models are seeded with a modest amount of manually labelled data. In practical applications involving meetings, such data can be obtained readily from brief round-table introductions. The second novel contribution relates to a incremental MAP adaptation procedure for efficient, on-line speaker modelling. When combined, these two developments provide an on-line diarization system which outperforms a baseline, off-line system by a significant margin. When configured appropriately, error rates may be low enough to support practical applications.

1. Introduction

Speaker diarization [1] aims to determine who spoke when in an audio stream. As per [2], the problem is formulated by:

$$\left(\tilde{S}, \tilde{G}, \tilde{\Delta}\right) = \underset{S,G,\Delta:S\in\Gamma(\Delta)}{\arg\max} P(S, G|\mathbf{O}), \quad (1)$$

$$\approx \underset{S,G,\Delta:S\in\Gamma(\Delta)}{\arg\max} P(\mathbf{O}|G, S), \quad (2)$$

where $\tilde{\Delta}$ represents an optimised speaker inventory, $\tilde{S}$ and $\tilde{G}$ represent an optimised speaker sequence and segmentation respectively, $\Gamma(\Delta)$ is the set of possible speaker sequences and $\mathbf{O}$ is the set of acoustic features. Speaker diarization has been investigated extensively in the contexts of broadcast news, lectures, phone conversations and meetings.

Historically, the state-of-the-art in speaker diarization has evolved around the development of off-line systems, i.e. where an audio stream is processed in its entirety before any segments are assigned speaker labels. Examples of such systems include [3–8]. Driven by the popularity of powerful, mobile smart devices, the need for real-time information extraction in human interaction, growing interest in the Internet of Things (IoT) and the proliferation of always listening sensors, on-line diarization has attracted increasing interest in recent years.

Due to their high computational complexity and latency, current state-of-the-art diarization techniques are not easily adapted to on-line processing. New, on-line approaches have thus been investigated, e.g. [9–11] and the first on-line system for meeting diarization [12]. Unfortunately, typical error rates are high – probably too high to support any practical applications. This has prompted us to re-evaluate the problem and to investigate alternative strategies.

This paper shows, unsurprisingly, that the bottleneck lies in the quantity of data used for speaker model initialisation. Two possible solutions to mitigate this bottleneck involve the relaxation of latency/on-line or supervision constraints. The former is at odds with the pursuit of on-line diarization. Whereas speaker diarization is traditionally an unsupervised problem, the work reported here investigates semi-supervised approaches.

While semi-supervised approaches have been reported previously for off-line diarization, this paper proposes a new, semi-supervised approach to on-line diarization. Based upon the approach in [12] and with the number of speakers assumed to be known a-priori, the new system uses a short duration of labelled speech for supervised speaker model initialisation. The remainder of the process remains entirely unsupervised. While knowing the number of speakers and the use of labelled data is also at odds with the traditional definition of diarization, many meeting scenarios involve a short round-table phase during which each speaker introduces themselves, data which may be used readily for initialisation.

While the manual labelling of such intervals is still an inconvenience, it is perhaps a price worth paying for the significant improvement in diarization performance. The goal of the work in this paper is thus to determine what duration of manually labelled speech is required in order to deliver satisfactory performance, here defined as that achievable with state-of-the-art off-line diarization. The second contribution of this paper relates to an incremental approach to on-line model adaptation which proves instrumental in delivering low error rates.

The remainder of this paper is organized as follows. Section 2 reviews prior work. Section 3 demonstrates the challenge faced in on-line diarization and justifies the need for relaxed supervision constraints. Section 4 describes the incremental model adaptation procedure and the new semi-supervised, on-line diarization system. Section 5 describes experimental work whereas conclusions and scope for future work are presented in Section 6.

2. Prior work

Although real-time diarization can be performed efficiently with the aid of multiple microphones and cameras [13, 14], diarization with a single microphone remains a challenge. On-

doi: 10.21437/Odyssey.2016-55

line diarization performance is then typically far from what can be achieved with off-line approaches. The past work is reviewed here.

Liu et al. [15] present an approach in the context of broadcast news diarization. Speech activity detection (SAD) is applied to identify speech segments which are clustered via one of two different algorithms in order to perform on-line diarization. The performance of the algorithms, involving leader-follower (LFC), dispersion-based (DSC) and combined clustering (hybrid speaker clustering, HSC), was evaluated on the NIST Hub4 1998 broadcast news database. Performance was assessed against a baseline off-line hierarchical clustering system. Average misclassification errors of 29.5% and 28.5% for the LFC and HSC algorithms and 35.5% for the DSC algorithm compared favourably to a baseline error of 31.5%.

Markov et al. [10, 11] investigated a more traditional approach using Gaussian mixture models (GMMs). Non-speech segments are discarded using a suitably trained GMM whereas diarization is performed upon the comparison of speech segments to a set of speaker models. New speaker models are introduced using an incremental expectation-maximisation (EM) algorithm. The system was assessed on a database of European Parliament plenary speeches, characterized by homogeneous and long speaker turns. A diarization error rate (DER) of 8% was reported.

A similar approach was reported by Geiger et al. [9] for broadcast news. Here, speaker models were learned through the maximum-a-posteriori adaptation (MAP) of a universal background model (UBM). The same UBM is used to control the attribution of speech segments to existing speakers and the addition of new speaker models. A DER of 39% was reported.

Vaquero et al. [16] present a hybrid system composed of off-line diarization [6] and on-line speaker identification. An initial off-line diarization stage is used to learn speaker models. An on-line speaker identification system is then used for subsequent diarization. Speaker models adaptation is performed in parallel. Performance is dependent on the latency and accuracy of the off-line process. A DER of 38% is reported for a set of 26 meetings from the NIST Rich Transcription (RT) evaluation corpora.

Oku et al. [17] report a low-latency, on-line speaker diarization system that makes use of phonetic information to estimate more discriminative speaker models. Phonetic boundaries are used as potential speaker turns. Acoustic features are clustered off-line into predefined acoustic classes. GMM speaker models have the same number of components as the number of acoustic classes. A traditional delta-BIC-like criterion is then used for speaker clustering and segmentation. Performance is assessed using Japanese TV talk shows where conversations are characterized by short speaker turns and only few silence intervals. A DER of 4% was reported.

The work in [12] proposed the first adaptive on-line speaker diarization system for NIST RT meeting data. Inspired by [9], speech segments of a fixed maximum duration obtained after speech activity detection (SAD) are classified against a set of speaker models learned by the MAP adaptation of a UBM. Despite the comparatively more difficult task, diarization error rates are similar to those reported in previous work, even if they are still high.

None of the above referenced work used any a priori information, information that may be readily harnessed in many practical applications in order to improve performance without significant impacts on convenience. This idea was investigated by Moraru et al. [18], albeit in the context of off-line diarization. Their work showed that even modest quantities of speaker training data bring significant performance improvements.

To the best of our knowledge, this paper presents the first work to assess what level of performance can be delivered in the context of on-line speaker diarization using similarly modest quantities of speaker training data. Fundamental to this is the second significant contribution which relates to an incremental MAP adaptation algorithm for the updating of speaker models during on-line processing.

3. Speaker modelling

Speaker diarization necessarily requires the learning of speaker models. In an on-line scenario, these are typically initialised using short speech segments. Diarization involves the comparison of similarly short, subsequent segments to the current inventory of speaker models $\tilde{\Delta}$ and possibly their consequent re-adaptation using steadily amassed data. While necessary to meet the requirements for on-line processing, the use of short segments for both operations also ensures inter-segment speaker homogeneity.

These two operations form the essential elements of *speaker verification*, namely speaker enrolment and testing. It is well known that the reliability of both depends fundamentally on data duration. The work presented in this section aims to examine the dependence of *speaker diarization* on segment duration and hence to illustrate the potential to improve on-line diarization performance. This examination is performed through strictly controlled speaker verification experiments which avoid complications associated with overlapping speakers [19] and compounding diarization nuances.

3.1. Databases and feature extraction

Automatic speaker verification (ASV) experiments are performed using four different datasets compiled from NIST Rich Transcription (RT) data. They are:

1. **UBM training (RTubm):** a set of 16 meeting shows from the NIST RT'04 evaluation;
2. **Development (RTdev):** a set of 15 meeting shows from the RT'05 and RT'06 evaluations, and
3. **Evaluation (RTeval):** a set of 8 meeting shows from the RT'07 and a set of 7 meetings from RT'09 evaluations.

where all data corresponds to the most challenging single distant microphone (SDM) condition of each standard evaluation subset. Acoustic signals are characterized by 19 Mel-frequency cepstral coefficients (MFCCs) augmented by energy, thereby obtaining feature vectors with a total of 20 coefficients computed every 10ms using a 20ms window.

3.2. Experiments

ASV is performed using a conventional GMM model with universal background model (GMM-UBM) system [20]. The UBM is trained using the meeting data from the RTubm dataset, with 10 iterations of expectation-maximisation (EM) and with 64 Gaussian components. Speaker models are derived from the UBM in the usual way using MAP adaptation with a relevance factor set to 10.

The speech data of all speakers with a floor time greater than 20 seconds are identified using the ground-truth references. The data for all other speakers are discarded. The first 10 seconds of speech of each speaker are set apart for model training

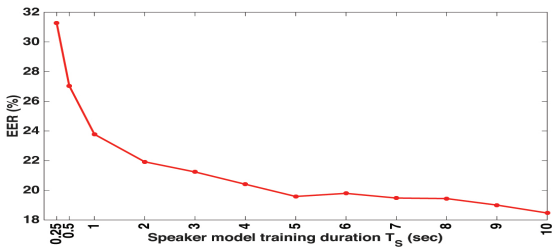


Figure 1: EER as a function of T_S, namely the quantity of data used to train the speaker models .

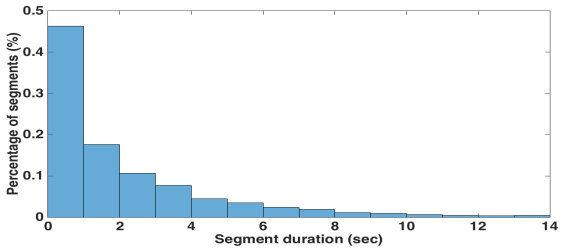


Figure 2: Speech segment duration distribution for the RTdev dataset.

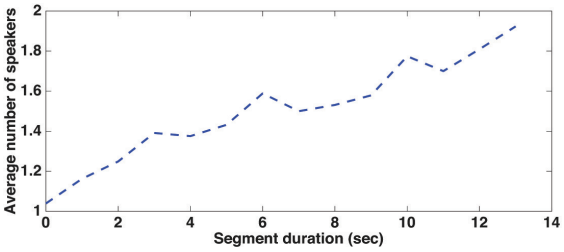


Figure 3: Average number of speakers as a function of the speech segment duration for the RTdev dataset.

while the remaining speech segments are used for testing. All speech segments are further divided into sub-segments of maximum duration T_S where $T_S = 0.25, 0.5, \ldots, 10$ seconds.

Training data, of identical duration T_S is randomly selected from the 10-second training segment, whereas testing is performed separately on every single, same-length sub-segment in the test data. Exhaustive testing is performed for all speakers; all test segments are compared to all speaker models. This equates to a large number of short-duration target and impostor trials from which ASV performance can be gauged in the usual way.

3.3. Results

ASV results, combined for RTdev and RTeval, are illustrated in Figure 1 in terms of the equal error rate (EER) as a function of T_S. Unsurprisingly, performance improves as T_S increases. Critically, with very low quantities of training and testing data less than 1 second in duration, the EER is extremely high. Lower EERs are observed for data quantities of 10 seconds. The elbow is around 5 seconds, where the EER is in the order of 20%. Even with a value of $T_S = 10$ seconds, the EER is perhaps still high for what is essentially same-session ASV. This is probably due to the fact that most speech segments are considerably shorter than the value of T_S.

Figure 2 illustrates the segment duration distribution for the RTdev dataset. The vast majority of segments are seen to be less than 5 seconds in duration. The use of longer segments in on-line speaker diarization applications also comes at the increased risk of speaker model impurities. Figure 3 illustrates the average number of speakers as a function of segment length. The plot shows that, beyond segment lengths of 5 seconds, a segment is more likely to contain 2 speakers than 1 speaker. Added to this, the use of longer segments would entail greater latency, which is at odds with the need for on-line diarization.

While admittedly trivial, this analysis shows that, in independence from overlapping speech and diarization nuances, the potential for successful on-line diarization is severely limited by the potential to acquire sufficient, speaker-homogeneous training and testing data. In summary, reliable decisions cannot be made when models are initialised on such short segments of speech. These observations call for an alternative approach to on-line diarization.

4. Semi-supervised on-line diarization

The use of larger segments for speaker model initialisation introduces latency and is at odds with the pursuit of on-line diarization. An alternative is needed which, in this paper, takes the form of relaxed supervision constraints. The penalty involves the use of short, manually labelled segments as seeds for speaker modelling. The target is to reach the same diarization performance obtained with an off-line system. The open question is what quantity of seed data is required to meet this objective?

While an admittedly trivial idea, the authors are not aware of any work in the open literature which reports such work in the case of *on-line* diarization. The novel contributions in the following are thus (i) the investigation of a new semi-supervised approach to on-line diarization and (ii) an incremental MAP adaptation procedure which improves significantly on our past results that used sequential MAP adaptation.

Before the on-line system is introduced, the differences between sequential and incremental MAP adaptation are first described. Even if MAP is a well known, standard algorithm, it serves here as the starting point. The conventional MAP adaptation algorithm is the first illustrated in Figure 4 which aims to illustrate the difference between the three MAP implementations described in the following.

4.1. Conventional maximum a-posteriori adaptation

The conventional MAP adaptation [20] algorithm is commonly used to adapt a UBM model, generally trained with an EM algorithm, towards a specific speaker. The algorithm calculates the posterior probability of each Gaussian component given a set of training observations, and can be applied to update the mean, covariance and weight parameters of the Gaussian components which have the highest posterior probabilities. In the case where speaker specific training data is scarce, then the MAP adaptation of a well trained UBM generally gives better results than speaker specific models learned directly by EM.

For a given speaker, let there be a sequence of D speech segments (D=4 in Figure 4) where each segment i is parametrised by a set of acoustic features $\boldsymbol{O}^{(i)} = \boldsymbol{o}_1, \ldots, \boldsymbol{o}_{M_i}$. As illustrated in Figure 4, conventional off-line MAP adaptation is performed using the UBM model λ_{UBM} and the pool of all D speaker segments. The sufficient statistics for the k-th

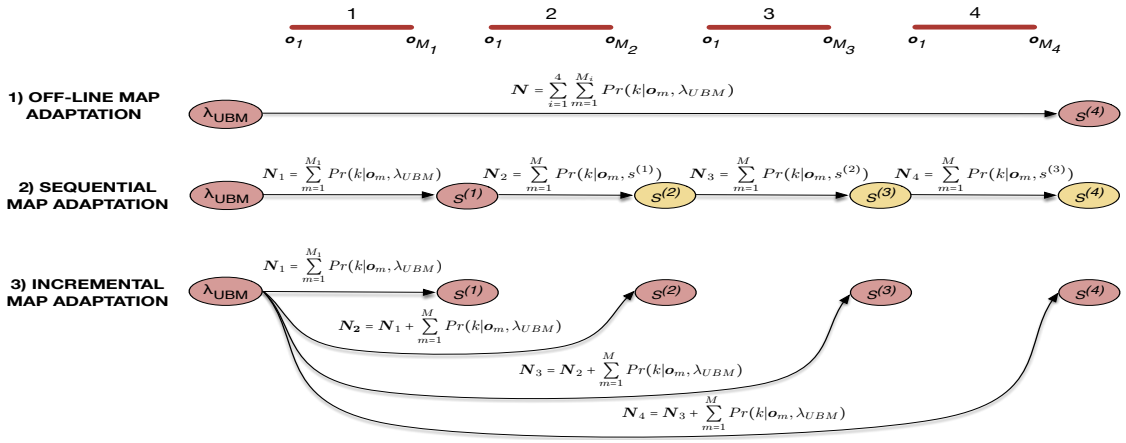


Figure 4: A comparison of off-line MAP adaptation, sequential MAP adaptation and incremental MAP adaptation for four speech segments from a particular speaker.

Gaussian component are obtained as follows:

$$
\begin{aligned}
N_k &= \sum_{i=1}^{D}\sum_{m=1}^{M_i} Pr(k|\boldsymbol{o}_m, \lambda_{UBM}) \\
\boldsymbol{F}_k &= \sum_{i=1}^{D}\sum_{m=1}^{M_i} Pr(k|\boldsymbol{o}_m, \lambda_{UBM})\boldsymbol{o}_m \\
\boldsymbol{S}_k &= \sum_{i=1}^{D}\sum_{m=1}^{M_i} Pr(k|\boldsymbol{o}_m, \lambda_{UBM})\boldsymbol{o}_m^2
\end{aligned} \tag{3}
$$

where $Pr(k|\boldsymbol{o}_m, \lambda_{UBM})$ represents the posterior probability of the k-th Gaussian component given the m-th observation $\boldsymbol{o}_m$. The MAP-adapted mean $\hat{\boldsymbol{\mu}}_k$, covariance $\hat{\boldsymbol{\sigma}}_k$ and weight $\hat{w}_k$ for the k-th Gaussian component are then given by:

$$
\begin{aligned}
\hat{w}_k &= \left(\alpha\frac{N_k}{\sum_{j=1}^{K} N_j} + (1-\alpha)w_k\right)\gamma \\
\hat{\boldsymbol{\mu}}_k &= \alpha\frac{\boldsymbol{F}_k}{N_k} + (1-\alpha)\boldsymbol{\mu}_k \\
\hat{\boldsymbol{\sigma}}_k &= \alpha\frac{\boldsymbol{S}_k}{N_k} + (1-\alpha)(\boldsymbol{\sigma}_k + \boldsymbol{\mu}_k^2) - \hat{\boldsymbol{\mu}}_k
\end{aligned} \tag{4}
$$

where γ is a normalization parameter such that $\sum_{k=1}^{K}\hat{w}_k = 1$ and where α is defined as:

$$
\alpha = \frac{N_k}{N_k + \tau} \tag{5}
$$

where τ is the pre-fixed scalar which regulates the relevance of the training data with respect to the UBM. The speaker model is then given by the tuple $s = (\hat{w}_k, \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\sigma}}_k)$.

4.2. Sequential MAP

Sequential MAP is the second algorithm illustrated in Figure 4 and the first approach suited to on-line processing used in our previous work [12]. Here, speaker models must be updated continuously as and when new speech segments are assigned to any one of the speaker models in the current speaker inventory.

An initial speaker model $s^{(1)}$ can be trained by calculating the sufficient statistics of the first speaker segment using the same UBM model λ_{UBM}. The sufficient statistics calculated for the k-th Gaussian components are obtained from the application of (3) while setting $D = 1$. The mean, variance and weights of the updated model $s^{(1)}$ are similarly obtained from (4). As soon as a new speaker segment is available, then speaker model $s^{(i)}$ can be updated using the sufficient statistics of the speaker segment $i+1$ and application of (3) with λ_{UBM} replaced by $s^{(i)}$:

$$
\begin{aligned}
N_{i+1} &= \sum_{m=1}^{M_{i+1}} Pr\left(k|\boldsymbol{o}_m, s^{(i)}\right) \\
\boldsymbol{F}_{i+1} &= \sum_{m=1}^{M_{i+1}} Pr\left(k|\boldsymbol{o}_m, s^{(i)}\right)\boldsymbol{o}_m \\
\boldsymbol{S}_{i+1} &= \sum_{m=1}^{M_{i+1}} Pr\left(k|\boldsymbol{o}_m, s^{(i)}\right)\boldsymbol{o}_m^2
\end{aligned} \tag{6}
$$

where subscripts k have been omitted for simplicity. The mean, variance and weights of the updated model $s^{(i+1)}$ are then obtained in the usual way using (4).

According to such a sequential procedure, the sufficient statistics N_{i+1}, $\boldsymbol{F}_{i+1}$ and $\boldsymbol{S}_{i+1}$ depend non-linearly on N_i, $\boldsymbol{F_i}$ and $\boldsymbol{S_i}$ in terms of Gaussian occupation probabilities. Accordingly, even given the same observations in the same segments, the speaker models obtained from the conventional, off-line and sequential MAP adaptation procedures are not the same.

4.3. Incremental MAP

The incremental procedure is the third algorithm illustrated in Figure 4. Here, the initial speaker model $s^{(1)}$ is obtained in the same way as with sequential MAP adaptation. In order to update the speaker model $s^{(i)}$, sufficient statistics for speaker segment $i+1$ are now always calculated with the original λ_{UBM}

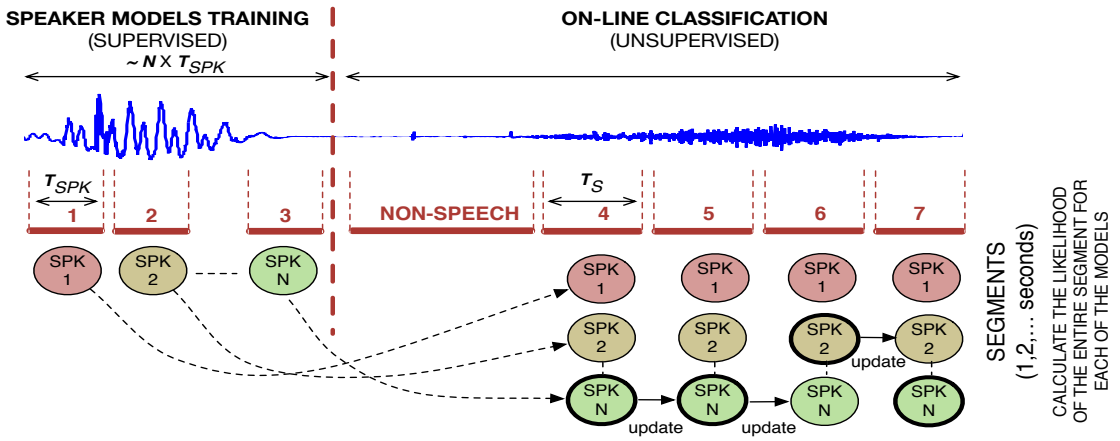


Figure 5: An illustration of the semi-supervised on-line speaker diarization system.

model and accumulated with sufficient statistics N_i, $\boldsymbol{F}_i$ and $\boldsymbol{S}_i$:

$$
\begin{aligned}
N_{i+1} &= N_i + \sum_{m=1}^{M_{i+1}} Pr\left(k|\boldsymbol{o}_m, \lambda_{UBM}\right) \\
\boldsymbol{F}_{i+1} &= \boldsymbol{F}_i + \sum_{m=1}^{M_{i+1}} Pr\left(k|\boldsymbol{o}_m, \lambda_{UBM}\right) \boldsymbol{o}_m \\
\boldsymbol{S}_{i+1} &= \boldsymbol{S}_i + \sum_{m=1}^{M_{i+1}} Pr\left(k|\boldsymbol{o}_m, \lambda_{UBM}\right) \boldsymbol{o}_m^2
\end{aligned} \tag{7}
$$

The mean, variance and weights of the updated model $s^{(i+1)}$ are then once more obtained according to (4). This procedure is linear and thus, given the same data, the incremental MAP procedure will produce the same models as the off-line procedure, while still being suited to on-line processing.

4.4. System overview

The on-line diarization system is illustrated in Fig. 5. It is based on the top-down or divisive hierarchical clustering approach to off-line diarization reported in [5] and the on-line diarization approach reported in [12].

The audio stream is parametrised as described in Section 2.1, thereby producing a stream of observations $\mathbf{O} = \mathbf{o}_1, \ldots, \mathbf{o}_T$. Critically, for any time $\tau \in 1, \ldots, T$ only those observations for $t < \tau$ are used for diarization. A brief round-table phase in which each speaker introduces himself is used to seed speaker models. The first T_{SPK} seconds of active speech for each speaker is set aside as seed training data.

An inventory $\tilde{\Delta}$ of speaker models s_j, where $j = 1, \ldots, M$, with M indicating the number of speakers in any particular meeting, is then trained using a certain duration of seed data T_{SPK} for each speaker. Speaker models are MAP adapted from the UBM using the seed data. For each speaker model s_j, the sufficient statistics $N_1^{(j)}$, $\boldsymbol{F}_1^{(j)}$ and $\boldsymbol{S}_1^{(j)}$ obtained during the MAP adaptation are stored in order to be used during the on-line diarization phase to update the speaker models. The resulting set of seed speaker models are then used to diarize the remaining speech segments in an unsupervised fashion.

4.5. On-line processing

Model-based speech/non-speech segmentation is adapted straightforwardly to on-line processing and is applied to remove non-speech segments. The remaining speech segments are then divided into smaller sub-segments whose duration is no longer than an a-priori fixed maximum duration T_S. On-line diarization is then applied in sequence to each sub-segment. The optimised speaker sequence $\tilde{S}$ and segmentation $\tilde{G}$ are obtained by assigning in sequence each segment i to one of the M speaker models according to:

$$
s_j = \underset{l \in (1,\ldots,M)}{\arg\max} \sum_{k=1}^{K} \mathcal{L}\left(\mathbf{o}_k | s_l\right) \tag{8}
$$

where $\mathbf{o}_k$ is the k-th acoustic feature in the segment i, K represents the number of acoustic features in the i-th segment and where $\mathcal{L}\left(\mathbf{o}_k | s_l\right)$ denotes the log-likelihood of the k-th feature in segment i given the speaker model s_l. The segment is then labelled according to the recognised speaker j as per (8). The updated speaker model s_j is obtained by either sequential or incremental MAP adaptation as described above.

5. Experiments

This section reports an evaluation of the new, semi-supervised speaker diarization system described above. The evaluation aims to determine what quantity of manually labelled seed data is needed to obtain the same performance as a state-of-the-art, entirely off-line system, the associated cost in terms of system latency and the benefit of incremental MAP adaptation.

5.1. Setup

Datasets, features and the UBM are exactly the same as those used for ASV experiments as reported in Section 3. Once again, all experiments concern the most challenging, single distant microphone (SDM) condition.

Since there is no round-table phase in the RT data, this component is simulated. Seed data is taken from wherever the first T_{SPK} seconds of speaker data are found. Since the majority of speakers speak for more than 2 minutes this has only a negligible bearing on the subsequent assessment of diarization performance. Overlapping speech is considered as non-speech and removed beforehand according to the ground-truth transcriptions.

T_{SPK}	3 sec.		5 sec.		7 sec.	
Algo. MAP	**Seq**	**Inc**	**Seq**	**Inc**	**Seq**	**Inc**
RTdev	24.7	21.3	21	18.1	20.5	16.5
RT07	19.1	17.3	17.5	14.6	13.6	13.3
RT09	23.7	18.2	17.6	16.2	21.2	16.2
Average	22.4	**18.9**	18.7	**16.3**	18.5	**15.3**

Seq = Sequential MAP; **Inc** = Incremental MAP

Table 1: A comparison of DER using sequential and incremental MAP algorithms. Results are reported for a segment duration / latency T_S of 3 seconds, three different datasets RTdev, RT07 and RT09 and for different durations T_{SPK} of training data.

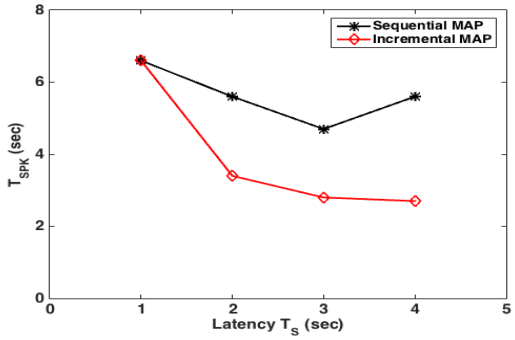


Figure 6: Speaker training data duration T_{SPK} against segment duration / latency T_S for the RT07 evaluation dataset using sequential and incremental MAP algorithms. All points correspond to a DER of 18 % (baseline, off-line performance).

Speaker models s_j are trained using quantities of labelled training data T_{SPK} of duration $1, \ldots, 39$ seconds. The maximum segment duration T_S is set to values of $0.25, 0.5, 1, \ldots, 4$ seconds. Larger values of T_{SPK} imply a greater inconvenience in manual training/enrolment. Larger values of T_S imply greater latency. Experiments were performed using both sequential and incremental MAP adaptation algorithms. Results for the RTdev and RTeval datasets are reported[1]. The baseline, off-line diarization system is the LIA-EURECOM top-down diarization system [21] with purification [5].

5.2. Results

Results in Figure 7 illustrate the variation in DER against the amount of speaker training data T_{SPK}. Left plots illustrate performance for sequential MAP adaptation whereas right plots correspond to incremental MAP adaptation. Results are illustrated independently for the RTdev (top), RT07 (middle) and RT09 (bottom) datasets. In each plot, different profiles illustrate performance for a range of segment durations / latencies T_S.

The first observation from Figure 7 indicates that the performance of the semi-supervised, on-line diarization system can surpass that of the baseline, off-line diarization system (illustrated with horizontal, dashed lines). In the case of sequential MAP adaptation this is achieved for the RTdev dataset, for instance, when speaker models are seeded with $T_{SPK} = 9$ seconds of training data when using a segment size / latency of $T_S = 4$ seconds. With the same segment size, the baseline performance for the RT07 and RT09 datasets is surpassed using as little as $T_{SPK} = 5$ and 3 seconds respectively.

In general, lower DERs are achieved with greater quantities of seed data, for instance a DER of 12.5% is achieved with $T_{SPK} = 9$ seconds of training data for the RT07 dataset and 15% with 17 seconds of training data for the RT09 dataset, both with latencies of $T_S = 3$ seconds.

Turning next to results for incremental MAP to the right in Figure 7, it is immediately evident that performance is significantly better than for sequential MAP. Here, the baseline, off-line diarization performance is surpassed with as little as $T_{SPK} = 5$ seconds of seed data for the RTdev dataset and $T_{SPK} = 3$ seconds in the case of both RT07 and RT09, all with a latency as low as $T_S = 2$ seconds. Once again, lower DERs are achieved with greater quantities of seed data, as low as 10% for the RT07 dataset and 12.5% for the RT09 dataset.

Table 1 summaries results across the three different datasets for T_{SPK}=3, 5, and 7 seconds of speaker training data and a fixed latency of $T_S = 3$ seconds. Results are illustrated for sequential and incremental MAP adaptation algorithms whereas average performance is illustrated in the bottom row. In all cases, incremental MAP adaptation delivers a lower DER.

Figure 6 plots the quantity of speaker training data T_{SPK} as a function of the latency T_S for the evaluation dataset RT07. All points correspond to a DER of 18% and thus show different configurations which achieve the same performance as the baseline, off-line diarization system. Plots are illustrated for both sequential and incremental MAP adaptation algorithms. In all cases, incremental MAP adaptation matches or betters baseline, off-line diarization performance with a lower amount of seed data or a lower latency than sequential MAP adaptation.

Finally, results presented in Figure 7 indicate that values of $T_S > 1$ seconds of latency are required for the best performance, no matter what is the value of T_{SPK}. Performance degrades universally for lower latencies. Crucially, for all datasets, DERs are equivalent or better than that of the baseline, off-line system when $T_S > 0.5$ seconds and when given sufficient training data T_{SPK}.

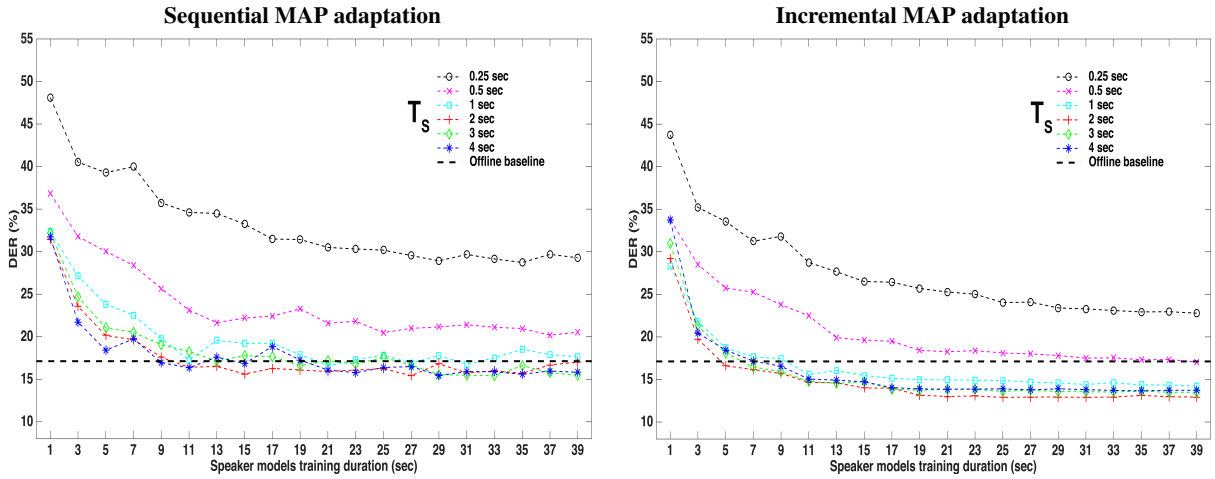
6. Conclusions and future work

This paper presents a semi-supervised on-line diarization system. The relaxation of supervision constraints overcomes the difficulty in initialising speaker models in an unsupervised fashion with small quantities of data; the use of longer segments would require effective speaker overlap detection and would also come at the expense of increased system latency.

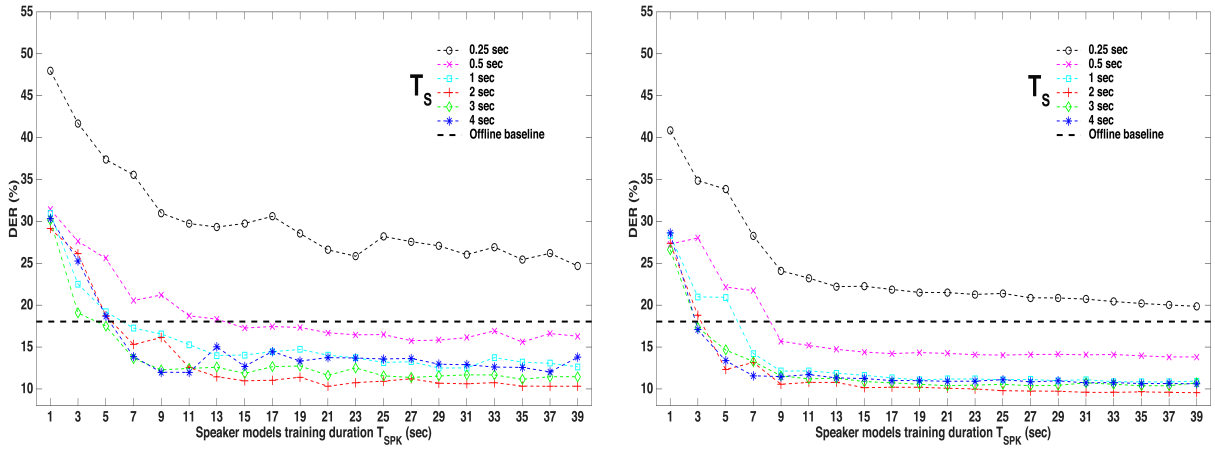
For the RT07 evaluation dataset, the work shows that such a system can outperform an off-line diarization system with just 3 seconds of speaker seed data and 3 seconds of latency with incremental MAP adaptation. By using greater quantities of seed data or by allowing greater latency, then a diarization error rate in the order of 10% can be achieved.

While these levels of performance may support practical applications, the need for supervised training remains an inconvenience. Reduced latency would also be welcome. If the inconvenience of a short, initial training phase proves acceptable, then this opens the potential for the application of either supervised or semi-supervised speaker discriminant feature transformations which may offer an opportunity for improved performance. This work could reduce the need for seed data, latency, or both. One avenue through which we are pursuing this objective involves phone adaptive training [22, 23].

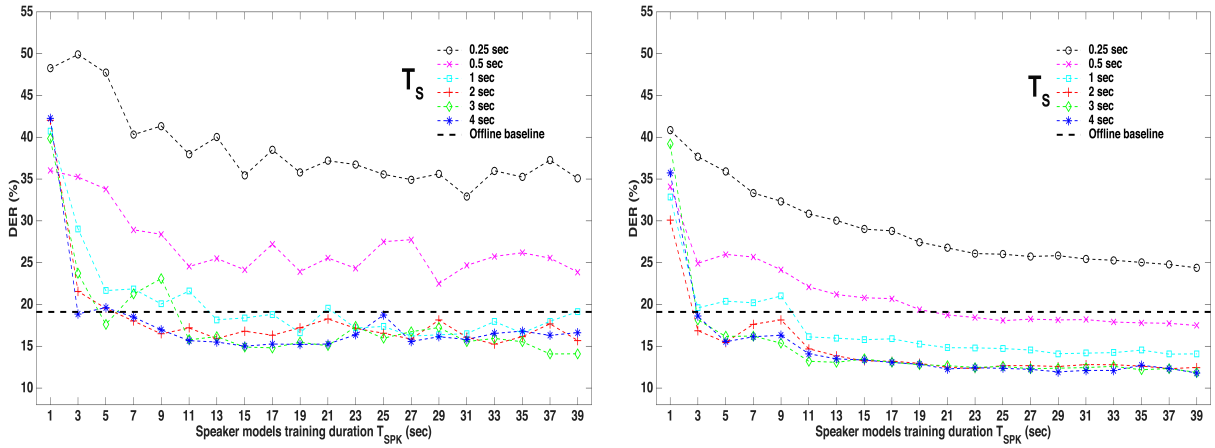
[1] In order to facilitate the comparison of results with the literature, results for the RTeval set are presented independently for the two subsets RT07 and RT09.



(a) Semi-supervised speaker diarization results for RTdev dataset



(b) Semi-supervised speaker diarization results for the RT07 evaluation dataset



(c) Semi-supervised speaker diarization results for the RT09 evaluation dataset

Figure 7: An illustration of DER for the semi-supervised on-line diarization system as a function of the speaker model training duration T_{SPK} and for different maximum segment durations / latency T_S. Results shown for the RTdev development, RT07 and RT09 evaluation datasets using sequential MAP adaptation (left) and incremental MAP adaptation (right). The horizontal, dashed line in each plot indicates the performance of the baseline, off-line diarization system.

7. References

[1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[2] N. Evans, S. Bozonnet, Dong Wang, C. Fredouille, and R. Troncy, "A comparative study of bottom-up and top-down approaches to speaker diarization," *IEEE Trans. Audio, Speech, Language Processing*, vol. 20, no. 2, pp. 382–392, 2012.

[3] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Multimodal Technologies for Perception of Humans*, pp. 509–519. 2008.

[4] N. W. D. Evans C. Fredouille, "The LIA RT'07 speaker diarization system," in *Lecture Notes on Computer Science, CLEAR 2007 and RT 2007, Multimodal Technologies for Perception of Humans*, 2008, vol. 4625/2008, pp. 520–532.

[5] S. Bozonnet, N. W. D. Evans, and C. Fredouille, "The LIA-Eurecom RT'09 speaker diarization system: enhancements in speaker modelling and cluster purification," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2010, pp. 4958–4961.

[6] G. Friedland, A. Janin, D. Imseng, X. Anguera, L. Gottlieb, M. Huijbregts, M.T. Knox, and O. Vinyals, "The ICSI RT'09 speaker diarization system," *IEEE Trans. Audio, Speech, Language Processing*, vol. 20, no. 2, pp. 371–381, 2012.

[7] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," in *Proc. Ann. Conf. International Speech Communication Association (INTERSPEECH)*, Aug. 2013.

[8] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE Trans. Audio, Speech, Language Processing*, vol. 17, no. 7, pp. 1382–1393, Sept 2009.

[9] J. T. Geiger, F. Wallhoff, and G. Rigoll, "GMM-UBM based open-set online speaker diarization.," in *Proc. Ann. Conf. International Speech Communication Association (INTERSPEECH)*, 2010, pp. 2330–2333.

[10] K. Markov and S. Nakamura, "Never-ending learning system for on-line speaker diarization," in *IEEE Workshop Automatic Speech Recognition Understanding (ASRU)*, Dec 2007, pp. 699–704.

[11] K. Markov and S. Nakamura, "Improved novelty detection for online GMM based speaker diarization," in *Proc. Ann. Conf. International Speech Communication Association (INTERSPEECH)*, 2008, pp. 363–366.

[12] G. Soldi, C. Beaugeant, and N. Evans, "Adaptive and online speaker diarization for meeting data," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Nice, France, 08 2015.

[13] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and J. Yamato, "Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera," *IEEE Trans. Audio, Speech, Language Processing*, vol. 20, no. 2, pp. 499–513, Feb 2012.

[14] S. Araki, T. Hori, M. Fujimoto, S. Watanabe, T. Yoshioka, T. Nakatani, and A. Nakamura, "Online meeting recognizer with multichannel speaker diarization," in *Conf. Signals, Systems and Computers (ASILOMAR)*, Nov 2010, pp. 1697–1701.

[15] D. Liu and F. Kubala, "Online speaker clustering," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, May 2004, vol. 1, pp. I – 333–6.

[16] C. Vaquero, O. Vinyals, and G. Friedland, "A hybrid approach to online speaker diarization," in *Proc. Ann. Conf. International Speech Communication Association (INTERSPEECH)*, 2010, pp. 2638–2641.

[17] T. Oku, S. Sato, A. Kobayashi, S. Homma, and T. Imai, "Low-latency speaker diarization based on bayesian information criterion with multiple phoneme classes," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, March 2012, pp. 4189–4192.

[18] D. Moraru, L. Besacier, and E. Castelli, "Using a priori information for speaker diarization," in *Odyssey - The Speaker and Language Recognition Workshop*, 2004.

[19] R. Vipperla, J. T. Geiger, S. Bozonnet, D. Wang, N. Evans, B. Schuller, and G. Rigoll, "Speech overlap detection and attribution using convolutive non-negative sparse coding," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2012.

[20] A. Reynolds, D., F. Quatieri, T., and B. Dunn, R., "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[21] C. Fredouille, S. Bozonnet, and N. Evans, "The LIA-EURECOM RT'09 speaker diarization system," in *RT09 NIST Rich Transcription Workshop*, Melbourne, Florida, USA, May 2009.

[22] S. Bozonnet, R. Vipperla, and N. Evans, "Phone adaptive training for speaker diarization," in *Proc. Ann. Conf. International Speech Communication Association (INTERSPEECH)*, 2012.

[23] G. Soldi, S. Bozonnet, Alegre F., C. Beaugeant, and N. Evans, "Short-duration speaker modelling with phone adaptive training," *Odyssey - The Speaker and Language Recognition Workshop*, 2014.